Chiem MAT 115 Homework 2

For the second homework assignment, we will explore the dataset `gapminder`, which contains data on health and income outcomes for 184 countries from 1960 to 2016. It is part of the `dslabs` package. You should write your answers in RMarkdown. The official due date for this assignment is **Sept. 16 at 2:30 PM**.

1. Take a look at the structure of the dataframe. How many cases are there? How many variables? What are the types of variables in the set? Are there any that seems mis-classified? If necessary, fix them.

A: 10545 rows, so there are 10545 cases. There are 3 different types variables: factor, int, and numeric. I think population may be mis-classified. It does not need to be numeric because none of the numbers are decimals.

```
library(dslabs)
gapminder$population <- as.integer(gapminder$population)
#class(gapminder$population)
#gapminder
#gapminder
```

2. Figure out what each column tells you.

A: In order from columns 1-9: name of country, year gathered, infant mortality rate, avg. life expectancy, fertility rates, population size, country's gdp, continent located in, and region located in

3. This is far too much data for one homework assignment. Subset the dataset to include only data from the year 2011. Be sure to give it an appropriate name. Once you do that, remove all the rows (countries) with missing GDP data (hint: you can use the `is.na` function).

```
#gapminder[gapminder$year == 2011 & !is.na(gapminder$gdp),]

gapminder_subset2011 <- data.frame(gapminder[gapminder$year == 2011 & !is.na(gapminder$gdp),])
```

4. If the dataset goes through 2016, why do you think I told you to use 2011?

A: The data after 2011 is lacking in some aspects especially their gdp. in 2016, most of the columns are missing data.

5. Which country has the highest GDP (definition of GDP: www.imf.org/en/Publications/fandd/issues/Series/Back-to-Basics/gross-domestic-product-GDP) in 2011? The lowest? What are the values?

A: Highest GDP = 1.174422e+13 (United States) and the Lowest = 77394709 (Kiribati)

```
#gapminder_subset2011$country[order(gapminder_subset2011)[1]]
```

6. Everyone (well, almost everyone) pays attention to the top and bottom, but what about the middle? I think it's useful to take a look at the typical value also. For this purpose, find the median GDP (very easy: just use the `median` function) and check if the median value actually equals any of the country values.

A: No there are not any that are equal to the median value

```
gapminder_subset2011[gapminder_subset2011$gdp == median(gapminder_subset2011$gdp),]
```

```
## [1] country          year             infant_mortality life_expectancy
## [5] fertility        population       gdp              continent
## [9] region
## <0 rows> (or 0-length row.names)
```

```
#gapminder[gapminder$year == 2011 & gapminder$gdp == median(temp_df$gdp),]
```

7. Perhaps raw GDP is not the best way to explore this dataset. Create a new variable, GDP per population, and add it to the dataframe. Then answer the same questions as before: which country has the greatest GDP per person? The least?

A: The least is at 109.2774 per capita (Democratic Republic of the Congo) and the most is 26928586464 per capita (Luxembourg).

```
gdp_per_capita <- gapminder_subset2011$gdp / gapminder_subset2011$population
#gdp_per_capita

gapminder_subset2011$gdp_per_capita <- gdp_per_capita
#gapminder_subset[order(gapminder_subset$gdp_per_capita, decreasing = TRUE)[1],]
```

8. Did you find anything striking or surprising in this dataset? (You can explore some more.) We didn't even look at the *geographical* or *health* components of the dataset. Provide a value or set of values that shows this striking aspect clearly.

A: I found surprising that Hong Kong had the highest life_expectancy in 2011. I originally expected it would be Japan (Although, Japan is third, with it being right behind Iceland by .30). In addition, it was also interesting to see the United States rank so low (36th) despite it being considered a "first world" or "developed" country. The US was ranked below some countries that some might consider "third world" or "developing countries" such as the Maldives, Chile, Saudi Arabia, and Costa Rica.

```
gapminder_subset2011[order(gapminder_subset2011$life_expectancy, decreasing = TRUE)[1:3],]
```

```
##               country year infant_mortality life_expectancy fertility
## 9509 Hong Kong, China 2011               NA           83.02      1.10
## 9511          Iceland 2011              1.8           82.90      2.11
## 9520            Japan 2011              2.3           82.60      1.39
##      population          gdp continent          region gdp_per_capita
## 9509    7044211 2.641367e+11      Asia    Eastern Asia       37496.99
## 9511     321030 1.110516e+10    Europe Northern Europe       34592.27
## 9520  127252900 5.058762e+12      Asia    Eastern Asia       39753.61
```

```r
gapminder_subset2011[order(gapminder_subset2011$life_expectancy, decreasing = TRUE)[seq(28, 36)],]
```

```
##            country year infant_mortality life_expectancy fertility population
## 9475    Costa Rica 2011              8.8            79.9      1.83    4600487
## 9481       Denmark 2011              3.2            79.9      1.76    5576577
## 9586      Slovenia 2011              2.6            79.9      1.49    2059023
## 9572         Qatar 2011              7.5            79.7      2.06    1905437
## 9541      Maldives 2011             10.1            79.6      2.31     338618
## 9526        Kuwait 2011              8.9            79.0      2.65    3239181
## 9469         Chile 2011              7.5            78.9      1.84   17201305
## 9579  Saudi Arabia 2011             14.2            78.9      2.76   28788438
## 9611 United States 2011              6.1            78.9      1.90  312390368
##               gdp continent           region gdp_per_capita
## 9475 2.536390e+10  Americas  Central America       5513.309
## 9481 1.710519e+11    Europe  Northern Europe      30673.274
## 9586 2.603711e+10    Europe  Southern Europe      12645.373
## 9572 6.760697e+10      Asia     Western Asia      35481.083
## 9541 1.311837e+09      Asia    Southern Asia       3874.091
## 9526 6.843929e+10      Asia     Western Asia      21128.580
## 9469 1.166315e+11  Americas    South America       6780.389
## 9579 2.784035e+11      Asia     Western Asia       9670.669
## 9611 1.174422e+13  Americas Northern America      37594.691
```

```r
# Costa Rica to US
```